Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.875
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Attention for Adaptive Convolution in Convolutional Neural Networks for Single Image Super-Resolution

Karam Park[a)] and Nam Ik Cho[a)‡]

Abstract

There have been many works on single image super-resolution (SISR) using convolutional neural networks (CNNs), researching on network architecture, loss function, applications, etc. However, very few have studied the modification or adaptation of the convolution operation, which is the fundamental element of CNN. In most CNN-based methods, the filter weights do not change at the inference phase, i.e., filter parameters are fixed regardless of the input and its regional characteristics. We note that this conventional approach is parameter-efficient but may not be optimal in performance due to its inflexibility to regionally different input statistics. To tackle this problem, we propose a novel convolution operation named Adaptive Convolution, which has content-specific characteristics. The proposed method adaptively adjusts filter weights according to the regional characteristics of the input with the help of an attention mechanism. We also introduce a kernel fragmentation method, which enables the efficient implementation of the Adaptive Convolution. We embed our new convolutional layer into several well-known SR networks and show that it enhances their performances while requiring a small number of additional parameters. Also, our method can be used along with other attentions that manipulate the features, further increasing the performance.

Keyword : Deep learning, Convolutional Neural Network, Single Image Super-resolution, Adaptive Convolution, Attention

a) Department of ECE, INMC, Seoul National University
‡ Corresponding Author : Nam Ik Cho
E-mail: nicho@snu.ac.kr
Tel: +82-880-8420
ORCID: https://orcid.org/0000-0001-5297-4649

※ This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023, in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (2021R1A2C2007220), and partially by Samsung Electronics Co., Ltd.
· Manuscript October 5, 2023; Revised December 14, 2023; Accepted December 14, 2023.

Ⅰ. Introduction

Single image super-resolution is a task that recovers a high-resolution (HR) output from its low-resolution (LR) counterpart by reconstructing lost information in the LR image. Due to its versatility and applicability, SISR is applied in diverse fields, such as medical imaging[32,35], satellite imaging[38], surveillance[33,46], and HDTV[9]. The SISR task is challenging due to its ill-posedness, meaning that

many different HR images can be SR results of the same LR input. To tackle this problem, various classical SR methods have been proposed, such as patch-based methods[4,8], statistics-based methods[23,42], sparse-coding-based methods[43,44], etc.

Recent developments in deep learning have promoted active research for deep learning-based SR methods. Dong et al.[6] proposed the first convolutional neural network (CNN) for the SR task, which outperforms classical SR algorithms[4,23,39,44]. By using only three convolutional layers, they demonstrated that CNNs have significant potential in the SR field. Afterward, numerous methods have been proposed to improve the performance of SR networks. The introduction of residual connections[13] enabled the SR network to have more layers, allowing the network to have larger receptive fields and achieve excellent performance. However, as the number of layers grows in the above networks, the computational cost also increases linearly whereas the performance gains are typically saturated. Moreover, simply increasing receptive fields does not guarantee the inclusion of more influential pixels for SR reconstruction, as mentioned in[10]. This implies that the unnecessary information also grows as the receptive field increases, impeding the performance increase.

The attention mechanism has been introduced to address this problem, allowing the network to focus on more helpful information. Channel attention (CA)[47] generates per-channel scaling factors to suppress unimportant channels and thus improve network performance. Non-local attention[5,30,48] extracts useful features from a long range of input pixels by capturing their distant dependencies, improving SR performance significantly at the cost of an enormous computational burden. As aforementioned, many studies have been conducted on applying attention for feature interdependencies, yet there are a few studies about using attention to filtering operations. The convolution operation is channel-specific and spatial-agnostic, i.e., different filter weights are applied to different channels, but the weights are fixed over the region. Notably, fixed weights are not ideal for handling the input with regionally varying statistics, which may lead to sub-optimal results[50].

In this paper, we propose Adaptive Convolution (AC) to tackle the above limitations, which can apply regionally different filters. We compute attention from the input and apply it to adjust the filter kernel. Hence, we can perform a different convolution on each spatial position. Specifically, we define the AC layer in which filter kernels are adjusted depending on input feature statistics. We also introduce a filter fragmentation method for the efficient implementation of our AC scheme. We define this scheme as the AC layer and apply it to several well-known SR networks. Extensive experiments show that our AC improves the performance in every case, often exceeding the performances of conventional CA. In addition, when we use CA and AC together, we can have more gains, showing that our AC can complementarily work with the conventional attention schemes.

Ⅱ. Related Works

1. Single Image Super-Resolution

Since the SRCNN first demonstrated that CNN has potential in the SR field, various deep learning-based SR methods have been proposed. Kim et al.[22] proposed VDSR, showing that increasing the depth effectively improves the performance. Shi et al.[34] proposed to apply the pixel-shuffle operation at the end of the network to improve previous inefficient upsampling methods. Ledig et al.[25] introduced residual connection into the SR network, showing that residual connection is an effective way to increase SR performance. The introduction of residual connection and the efficient upsampling operation[34] resulted in significant performance improvement through depth-wise expansion[28].

In addition, various network designs have been suggested further to improve the performance or efficiency of the SR network. Recursive network design[21,36,37] employs the same modules recursively to learn LR-HR mapping without increasing parameters. Multi-path network design[11,27] handles features with multiple paths to achieve better representational ability. Dense connection designs[12,49] reuse features from preceding layers[16] to enhance the reconstruction performance. Complex residual connection designs[14,31] have been proposed to explore more efficient residual blocks. As mentioned above, various structural developments have been attempted, yet few studies have considered spatially-adaptive convolution, which is our main objective.

2. Attention Mechanism for SR

The attention mechanism enables the network to focus more on informative relations or locations. Channel attention (CA) enhances the network representations by considering the interdependencies of channels. With global average pooling and two dense layers, CA produces channel-wise scaling factors to suppress less important channels. Based on the Squeeze-and-Excitation block[15], Zhang et al.[47] proposed RCAN, adopting CA into the SR network. Non-local attention helps the network consider relationships between features in spatially distant locations.

Non-local attention captures all possible pair-wise feature interdependencies, and it was also adopted to SR networks to overcome the limitation of local receptive fields of the plain CNN[5,30,48]. Unlike these attention methods that find spatial and channel correlation to control their contributions, we apply the attention to control the filter kernel to apply spatially adaptive filtering.

3. Dynamic Kernel

The traditional convolution operation in CNNs is content-independent, i.e. the same filter is used for all positions of the input, which is suboptimal for the input with spatially varying properties. In order to deal with this shortcoming, Xu et al.[19] introduced a dynamic filtering scheme, which has a filter-generating network and dynamic filtering layer. Zhou et al.[50] proposed computationally efficient Decoupled Dynamic Filter Networks by decoupling a depth-wise dynamic filter into spatial- and channel-dynamic filters. Li et al.[26] proposed an efficient and effective operator Involution, which inverts the spatial-independent and channel-specific characteristics of convolution.

Commonly, these dynamic convolution methods utilize a filter-generating module to directly generate kernels for each pixel. This generally requires many computations and excessive memories for storing different filter coefficients for the whole number of pixels. Hence, despite the advantage of improving feature representation, it is challenging

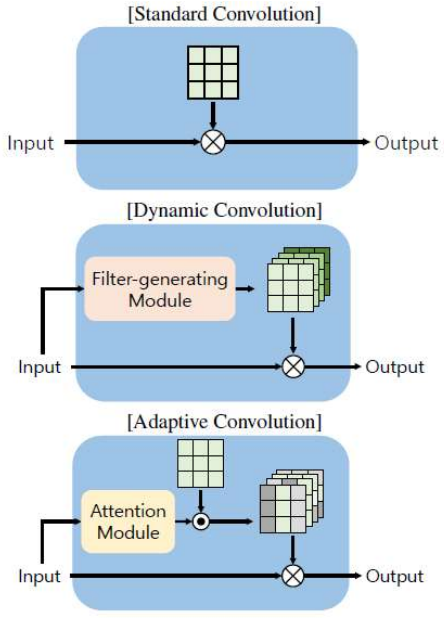

Fig. 1. Comparison between Convolution operations. (a) Standard Convolution applies a static filter globally. (b) Dynamic Convolution generates filters for each location via Filter-generating Module. (c) Proposed Adaptive Convolution calculates the attention map via the Attention Module and generates per-pixel filters using the attention map and static filter

to apply conventional dynamic filtering methods for the SR. To resolve this problem, we introduce an attention mechanism for adaptive filtering and its efficient computation by our proposed filter fragmentation. Unlike previous methods, we use an attention module that predicts regionally appropriate filters. Precisely, we adaptively scale parts of the static filters according to the attention scaling factors. The proposed filter fragmentation method simplifies the dynamic filtering process into element-wise multiplication of attention maps and features, which is described in Section Ⅲ.2. The difference between standard convolution, dynamic convolution, and the proposed AC is illustrated in Figure 1.

Ⅲ. Adaptive Convolution

The classic convolution operation in CNNs is to apply a static filter for all pixels of the input. Using a static filter is efficient regarding the computation time and structural modularity, but it may yield sub-optimal results for the inputs with varying statistics[50]. In this section, we introduce the AC that applies regionally different weights to cope with spatially varying properties of inputs.

1. Concept of Adaptive Convolution

To define the AC, we first consider the standard convolution operation defined as

$$X'(i) = \sum_{j \in \Omega(i)} W(p(i,j)) X(j) \quad + b \tag{1}$$

where $X'(i) \in \mathrm{R}^{Co}$ denotes the output feature vector at the i-th pixel, $X(j) \in \mathrm{R}^{Ci}$ denotes the input feature at the j-th pixel, $\Omega(i)$ represents the pixels within the range of the convolution window (K×K) around the i-th pixel, $W \in \mathrm{R}^{K \times K \times Ci \times Co}$ denotes a convolutional kernel with the size of K, $p(i,j)$ denotes the offset defined by the distance and direction between the i-th and the j-th pixel, and b denotes the bias. As shown above, the fixed filter W is shared across the input feature vectors, regardless of their content. To assign spatially adaptive characteristics to the convolution operation, we apply the attention mechanism to suppress or enhance input at a specific location based on the pixel content. The new operation for this purpose, named AC, is defined as:

$$X'(i) = \sum_{j \in \Omega(i)} W'(p(i,j),i) X(j) + b \tag{2}$$

$$W'(p(i,j),i) = W(p(i,j)) A(p(i,j), X(i)) \tag{3}$$

where $W' \in \mathrm{R}^{K \times K \times Ci \times Co \times H \times W}$ is the weight of the AC, and $A \in \mathrm{R}^{K \times K \times H \times W}$ is the attention scaling factors, which is conditioned on the input feature vector at the i-th pixel $X(i)$. The generalized AC described in above equations is illustrated in Figure 2.

2. Implementation of Adaptive Convolution

For implementing spatially varying convolution to the

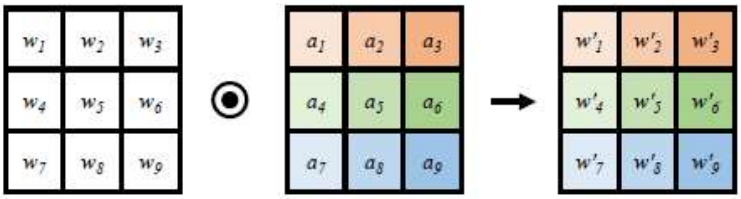

Fig. 2. Filter generation process with the kernel size of 3. $w_n \in \mathrm{R}^{Cin \times Cout}$ denotes the weight vector of the filter at the offset n, ⊙ represents element-wise multiplication, and $a_n \in \mathrm{R}^{H \times W}$ represents its corresponding attention scaling factor. For the convenience of visualization, channel and spatial dimensions are omitted

image of size H×W, it may seem that the dimension of the filter must be increased from $\mathrm{R}^{K\times K\times Ci\times Co}$ (memory for storing the static filter) to $\mathrm{R}^{K\times K\times Ci\times Co\times H\times W}$, which unduly increases the system complexity. One of the efficient methods to implement this idea is to split the filter and apply attention to the features created by fragment filters rather than directly to the whole filter. From Equations (2) and (3), this method can be described as:

$$\begin{aligned} X'(i) &= \sum_{j\in\Omega(i)} W(p(i,j))A(p(i,j),X(i))X(j)+b \\ &= \sum_{j\in\Omega(i)} a_{p(i,j)}(i)w_{p(i,j)}X(j)+b \end{aligned} \tag{4}$$

where $w_{p(i,j)}\in\mathrm{R}^{1\times 1\times Ci\times Co}$ is the fragment filter from the whole filter W, and $a_{p(i,j)}(i)\in\mathrm{R}^{1\times 1}$ represents the attention scaling factor for the fragment filter $w_{p(i,j)}$ at the i-th pixel. The attention scaling factors from the attention module are computed as:

$$[a_1,\ldots,a_n]=M(X) \tag{5}$$

where $a_n\in\mathrm{R}^{H\times W}$ represents the attention scaling factor of the offset n, and M refers to the attention module. The implementation of generalized AC is illustrated in Figure 3.

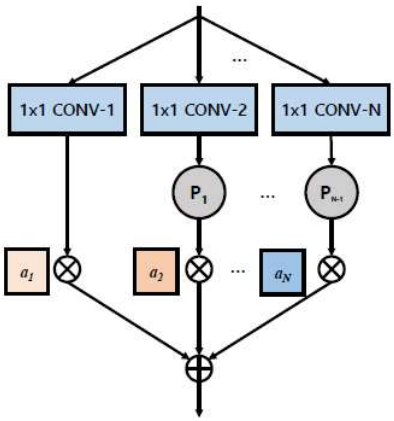

Fig. 3. Implementation method of generalized Adaptive Convolution. 1×1 CONV-n refers to fragment filter w_n, P_n represents padding operation corresponding to offset n, and a_n is an attention scaling factor of its counterpart fragment filter w_n

3. Design of Adaptive Convolution

Following the method in Section 3.2, the 3×3 general AC layer can be implemented with nine 1×1 convolution filters. In addition, we design the AC layer that is more suitable for the SR task. Inspired by the Sobel operator [20], where the filter is divided into three parts, we propose a method of dividing a 3×3 convolution filter into three 3×1 and three 1×3 fragment filters. For this, attention scaling factors from the attention module share the same value row-wise $A(\bullet,\bullet)\in\mathrm{R}^{3\times 1\times H\times W}$ or column-wise $A(\bullet,\bullet)\in\mathrm{R}^{1\times 3\times H\times W}$, which is illustrated in Figure 4.

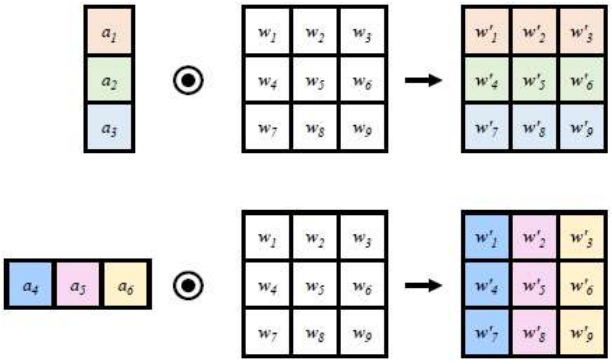

Fig. 4. Illustration of the filter generation process of Adaptive Convolution with the fragment 3×1 and 1×3 filters. (Top) 3×1 fragment filter shares attention scaling factor row-wise. (Bottom) 1×3 fragment filter shares attention scaling factor column-wise

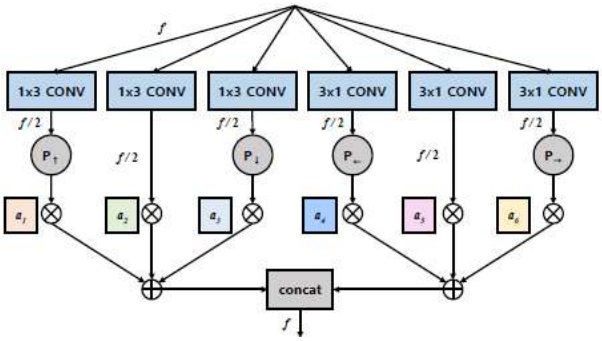

Fig. 5. Implementation of the Adaptive Convolution with the fragment 1×3 and 3×1 filters. 1×3 and 3×1 CONV refer to fragment filters, $\mathrm{P}_{(\bullet)}$ represents padding operation corresponding to its offset, f means the number of channels, and a_n is an attention scaling factor of its corresponding fragment filter

To obtain balanced feature representations, we acquire half of the feature channels from 1×3 filters and the other half from 3×1 filters. The above-mentioned process is depicted in Figure 5.

For the SR task, experiments in Section Ⅳ.3 show that the AC layer implemented with 3×1 and 1×3 is a better choice than several 1×1 filters. This is because edge-related information is important in the SR task, and it is more advantageous to consider multiple adjacent pixels together than a single pixel to collect edge-related information. Hence, in the rest of this paper, we consider AC as a 3×3 convolution layer implemented with 3×1 and 1×3 fragment filters.

4. Application of Adaptive Convolution

The residual block (Fig. 6a) is a widely adopted structure for SR networks. There are three possible ways to apply the AC, as shown in Figs. 6b, 6c and 6d. The ablation studies in Section Ⅳ.3 show that replacing the first convolution layer with the AC (see Fig. 6b) yields the best results for the SR. This indicates that convolution layers in residual blocks have different purposes, and it is important for the first convolution layer to receive edge-related information selectively. A detailed explanation is provided in Section Ⅳ.3.

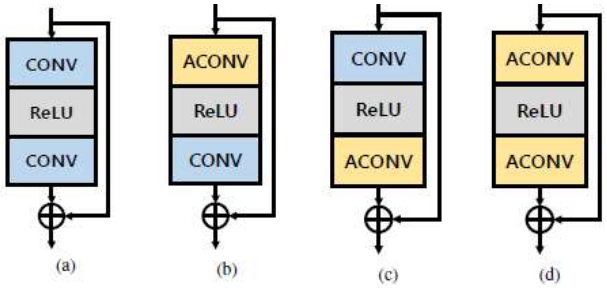

Fig. 6. (a) Residual block, (b),(c),(d) Possible ways of applying the Adaptive Convolution to Residual block

Ⅳ. Experimental Results

1. Implementation Details

As the attention module for controlling the AC, a simple structure consisting of a 1×1 convolution layer and a sigmoid activation layer is used. For a fair comparison, we

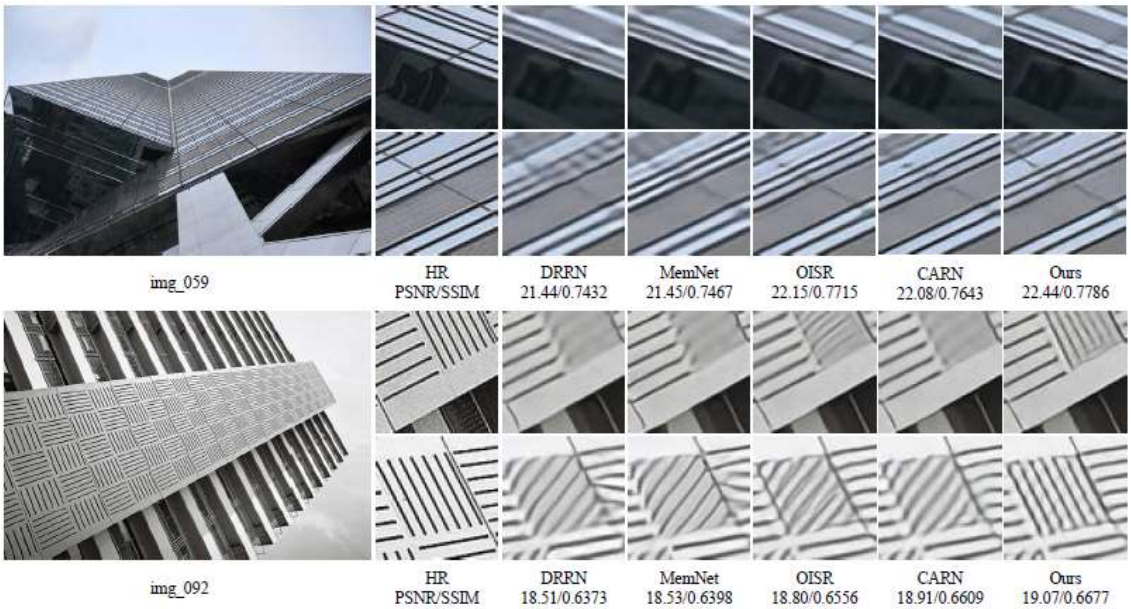

Fig. 7. Qualitative Comparisons with state-of-the-art methods on the Urban100 (×4) datasets. Our results are collected from EDSR-b16c64-A, which is the EDSR baseline with the Adaptive Convolution

implement and train all compared models in the same environment. Specifically, all models are trained with 800 train images from DIV2K dataset[1], following most of the previous works. Also, we augment data with combinations of flips and rotations and use 48×48 sized RGB patches. For the evaluation, four benchmark datasets are selected including Set5[3], Set14[45], BSD100[29], and Urban100[17]. For the generation of LR images, we used bicubic downsampling, following previous methods. For quantitative comparison, we use the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM)[41] on the luminance (Y) channel of test images. We set the minibatch size as 16. For training, we use ADAM optimizer[24] with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4}, and the decaying factor is set to 0.85 for every 2×10^5 iterations. We use the L1 loss as a loss function. All results are trained and evaluated on the NVIDIA TITAN XP GPU device.

2. Effectiveness of the Adaptive Convolution in SR networks

To show the effectiveness of our AC as a new convolution layer, we first apply the AC to the simple EDSR

Table 1. Quantitative comparisons of EDSR baseline added with our AC and previous networks with similar numbers of parameters.

Scale	Method	Param	Set5		Set14		BSD100		Urban100	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	FSRCNN	0.01M	37.00	0.9558	32.63	0.9088	31.53	0.8920	29.88	0.9020
	DRRN	0.30M	37.74	0.9591	33.23	0.9136	32.05	0.8973	31.23	0.9188
	MemNet	0.68M	37.78	0.9597	33.28	0.9143	32.08	0.8978	31.31	0.9195
	OISR-LF-s	1.37M	38.02	0.9605	33.62	0.9178	32.20	0.9000	32.21	0.9290
	CARN	1.59M	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256
	EDSR-b16c64	1.37M	38.02	0.9607	33.60	0.9180	32.20	0.9004	32.15	0.9291
	EDSR-b16c64-A	1.38M	38.08	0.9610	33.74	0.9191	32.22	0.9006	32.26	0.9299
	OISR-LF	4.97M	38.12	0.9609	33.78	0.9196	32.26	0.9007	32.52	0.9320
	MSRN	5.89M	38.08	0.9605	33.74	0.9170	32.23	0.9013	32.22	0.9326
	EDSR-b24c96	4.41M	38.10	0.9610	33.77	0.9194	32.26	0.9011	32.50	0.9321
	EDSR-B24C96-A	4.42M	38.16	0.9611	33.80	0.9190	32.28	0.9013	32.56	0.9326
×3	FSRCNN	0.01M	33.16	0.9140	29.43	0.8242	28.53	0.7910	26.43	0.8080
	DRRN	0.30M	34.03	0.9244	29.96	0.8349	28.95	0.8004	27.53	0.8378
	MemNet	0.68M	34.09	0.9248	30.00	0.8385	28.96	0.8001	27.56	0.8376
	OISR-LF-s	1.55M	34.39	0.9272	30.35	0.8426	29.11	0.8058	28.24	0.8544
	CARN	1.59M	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493
	EDSR-b16c64-A	1.56M	34.50	0.9276	30.40	0.8442	29.15	0.8072	28.32	0.8563
	OISR-LF	5.64M	34.56	0.9284	30.46	0.8450	29.20	0.8077	28.56	0.8606
	MSRN	6.08M	34.38	0.9262	30.34	0.8395	29.08	0.8041	28.08	0.8554
	EDSR-b24c96-A	4.84M	34.55	0.9281	30.47	0.8453	29.21	0.8089	28.56	0.8610
×4	FSRCNN	0.01M	30.48	0.8628	27.49	0.7503	26.90	0.7101	24.52	0.7221
	DRRN	0.30M	31.68	0.8888	28.21	0.7720	27.38	0.7284	25.44	0.7638
	MemNet	0.68M	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630
	OISR-LF-s	1.52M	32.14	0.8947	28.63	0.7819	27.60	0.7369	26.17	0.7888
	CARN	1.59M	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837
	EDSR-b16c64-A	1.52M	32.24	0.8946	28.65	0.7838	27.62	0.7386	26.20	0.7904
	OISR-LF	5.50M	32.33	0.8968	28.73	0.7845	27.66	0.7389	26.38	0.7953
	MSRN	6.33M	32.07	0.8903	28.60	0.7751	27.52	0.7273	26.04	0.7896
	EDSR-b24c96-A	4.75M	32.36	0.8960	28.75	0.7863	27.68	0.7410	26.44	0.7978

baseline[28] and see whether they improve the performance. Then, we generalize the claim by applying the AC to other well-known SR methods by showing their improvements.

We prepare several variants of EDSR baseline and add the AC to each of them, and compare with several well-known SR networks having similar numbers of parameters in Table 1. Note that we do not claim that the EDSR+AC works better than recent state-of-the-art SR networks in this table, but show how much the AC improves the baseline and its performances compared to similar networks. Our main claim is that our AC can improve many other SR networks, including more recent methods than the ones in Table 1, which will be addressed later in Section Ⅳ.3 and Table 5.

Regarding the notations of EDSR baseline in Table 1, the numbers of residual blocks and channels are written after b and c, respectively. When a model's first layer of the residual block is changed to the AC layer, it is denoted with the suffix "-A." The table shows that our EDSR-b16c64-A models perform comparably to other similar-complexity methods. Also, Figure 7 shows that our method provides visually better results. Regarding the complexity, compared to one convolution layer with 64 channels having around 37K of parameters, our parameter overhead due to the attention module is only 0.39K per residual block.

3. Ablation Studies

In this section, we investigate the effect of the proposed method and compare its performance with possible alternatives mentioned in Sections Ⅲ.2 and Ⅲ.3. All ablation studies are performed using the EDSR baseline (EDSR-b16c64) as a base network model, on the benchmark datasets (×2).

First, we investigate the best position of the AC layer in a residual block. As described in Figs. 6b, 6c and 6d, we select these three methods as candidates. The reconstruction performances of each candidate are shown in Table 2. In Table 2, we denote the convolution layer, ReLU activation layer, and the proposed AC layer as C, R, and A, respectively. From the table, we observe that the proposed A-R-C method shows significant performance improvement compared to the C-R-C baseline. Note that our method has the same convolutional operation as the baseline, except that the attention module suppresses uninformative pixels from being received. Compared to other candidates, the proposed A-R-C residual block shows slightly better performance. Interestingly, the A-R-A residual block shows mediocre performance even with more AC layers. This result indicates that two convolution layers in the residual block may have different purposes, and constraint operation of both layers with attention may hinder the function of the residual block in the SR network.

Table 2. Ablation studies on the choice of the position in the residual block. C represents the convolution layer, R is ReLU activation Layer, and A refers to the Adaptive Convolution layer.

Block	Param	Set5	Set14	BSD100	Urban100
C-R-C	1.37M	38.02	33.60	32.20	32.15
A-R-C	1.38M	38.08	33.74	32.22	32.26
C-R-A	1.38M	38.07	33.70	32.22	32.26
A-R-A	1.38M	38.04	33.70	32.23	32.26

Next, we analyze the performance of the model with the AC using different fragment filter settings. The comparison results are shown in Table 3, where 9×(1×1) is the generalized AC filter in Section 3.1, and 3×(1×3, 3×1)

Table 3. Ablation studies on the choice of fragment filter. The first convolution layer of the residual block is altered with the AC layer with the other fragment filters, where we use EDSR-baseline (EDSR-b16c64) as a baseline model

Filter	Param	Set5	Set14	BSD100	Urban100
baseline	1.37M	38.02	33.60	32.20	32.15
9×(1×1)	1.38M	38.05	33.73	32.21	32.21
2×(3×3)	1.38M	38.05	33.69	32.20	32.16
3×(1×3,3×1)	1.38M	38.08	33.74	32.22	32.26

is the proposed method in Section 3.3. In addition to this, we add 2×(3×3) with the exactly same number of parameters as the proposed method to compare the effect of increasing parameters on the performance. Strictly speaking, 2×(3×3) is not the AC, but it just has two 3×3 branches and an attention mechanism similar to the proposed method. For 2×(3×3), three attention scaling factors are applied to each 3×3 branch. As shown in Table 3, with a slight increase of parameters, our two methods show significant performance improvements compared to the baseline. Even though 9×(1×1) has more options in suppressing the filter, 3×(1×3, 3×1) performs better. This implies that suppressing the filter row-wise or column-wise is better than pixel-wise for edge finding, like the Sobel operator. By comparing the proposed method with 2×(3×3), we demonstrate that simply increasing the parameter does not significantly improve the performance. We describe in detail how the proposed filter attention works in Section Ⅳ.4.

Then, we compare our method with the most widely used attention method for SR, CA. In addition, we apply both CA and the proposed approach to the baseline and compare its performance, denoted as "+ AC & CA." Table 4 compares how our AC and the conventional CA are affecting the performance of the baseline network. From the 2nd and 3rd rows, we can see that our AC is slightly better than the CA while requiring slightly fewer parameters. Also, the 4th row shows that using both AC and CA does not improve the results in the case of Set5, Set14, and BSD100, but improves the performance for Urban100. We conjecture that two different attention schemes work complementarily for the Urban100, which contains images with high-frequency structures that are difficult to restore correctly. Considering that most of the previous attention methods also scale the features resulting from the convolution layer, we believe our method can be concurrently used with other attention methods as well.

Finally, we apply our method to other SR network models and compare their performances. To show that the proposed method has generality, we apply our method to IMDN[18] and A2F[40], which have complex structures that do not resemble EDSR structures. Table 5 compares the performance of networks with and without AC. All models in Table 5 are trained from scratch with the settings in Section Ⅳ.1. With a small parameter overhead, our method improves the reconstruction performance of both models. Note that IMDN[18] and A2F[40] have complicated handcrafted structures with CA-variants. Through ablation studies, we show that the proposed method is applicable to various structures other than simple EDSR-like structures.

Table 4. Comparison with the widely used channel attention (CA) method. We also test the simultaneous use of the proposed method with CA.

Method	Param	Set5	Set14	BSD100	Urban100
baseline	1.37M	38.02	33.60	32.20	32.15
+ AC	1.38M	38.08	33.74	32.22	32.26
+ CA	1.38M	38.04	33.70	32.22	32.24
+ AC & CA	1.39M	38.09	33.72	32.22	32.35

Table 5. Ablation studies on different models. The first convolution layer of the residual block is altered with the Adaptive Convolution layer. All models are trained from scratch by ourselves. Models with the Adaptive Convolution method are denoted with the suffix "-A."

Method	Param	Set5	Set14	BSD100	Urban100
IMDN	0.69M	38.03	33.56	32.15	32.09
IMDN-A	0.70M	38.06	33.69	32.18	32.20
A2F-L	1.36M	38.10	33.75	32.23	32.43
A2F-L-A	1.37M	38.10	33.78	32.24	32.50

4. Model Analysis

In this section, we investigate how the proposed method works in the SR network by examining the attention maps used in each layer. As shown in Figure 8, we compared attention maps for 1×3 and 3×1 fragment filters at the 1st, 4th, 8th, 12th, and the last residual blocks. To visualize attention maps, we use attention values from three channels

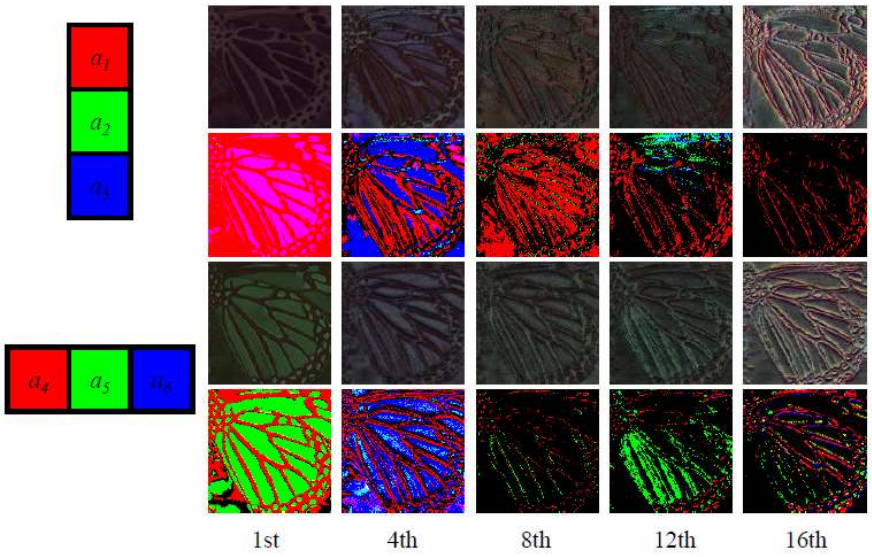

Fig. 8. Visualized attention maps of the EDSR-b16c64-A (×2) in the 1st, 4th, 8th, 12th, and the last residual blocks. The color expressed for each pixel indicates the characteristics of the filter at the position. The color difference between pixels means that different convolution operations are used at each location

as RGB channel values. For visibility, we also compare attention maps normalized with their maximum value.

First, we compare the patterns of the attention map responses. Regardless of the filter shape, attention maps show a sparse response as the layer deepens. This is consistent with the result of previous works showing a sparse response near edges as the SR network deepens. As the depth increases, we observe that different patterns of responses appear depending on the shape of the fragment filters. For example, attention maps for the 1×3 filters (first and second row) tend to respond near vertical edges as the layer deepens. This is because the shape of the filter 1×3 is advantageous to grasp the vertical edge, and the attention mechanism helps the filter to focus more on the vertical edges.

Next, we investigate the response of the attention map at each pixel. It is noticeable that various colors are observed in the same activation map. These results imply that various filters are utilized in the same layer, and these adaptive filters, conditioned on the pixel content, contribute to performance improvement as intended. Intriguingly, in the shallow layers, it is observed that a part of the filter is used to distinguish image parts with different characteristics. Considering that the proposed method is designed for discriminating edges better, the network seems to have utilized the proposed method by adapting to the shallow layer environment. We also observe that various colors appear in the attention map as layers get deeper. This indicates that the deeper the layer, the more various filters are required for better feature representation.

Ⅴ. Conclusion

In this paper, we have introduced Adaptive Convolution, a novel convolution operation that utilizes an attention mechanism to adjust the filter weights regarding the regionally varying context of images. The proposed Adaptive Convolution, whose weights are controlled by attention, improves the feature representation by preventing less-in-

formative pixels from being received. Through extensive experiments, we have demonstrated that the proposed method utilizes various filters at each location and improves the network performance with a small parameter overhead. In future works, we plan to apply the proposed method to other image restoration tasks, enhance the performance through structural improvement, and explore its more efficient implementation methods.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 126 – 135, 2017.
doi: https://doi.org/10.1109/cvprw.2017.150

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 252 – 268, 2018.
doi: https://doi.org/10.1007/978-3-030-01249-6_16

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.
doi: https://doi.org/10.5244/c.26.135

[4] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 1, pages I – I. IEEE, 2004.
doi: https://doi.org/10.1109/cvpr.2004.1315043

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11065 – 11074, 2019.
doi: https://doi.org/10.1109/cvpr.2019.01132

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2):295 – 307, 2015.
doi: https://doi.org/10.1109/tpami.2015.2439281

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In European conference on computer vision, pages 391 – 407. Springer, 2016.
doi: https://doi.org/10.1007/978-3-319-46475-6_25

[8] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. IEEE Computer graphics and Applications, 22(2):56 – 65, 2002.
doi: https://doi.org/10.1109/38.988747

[9] Tomio Goto, Takafumi Fukuoka, Fumiya Nagashima, Satoshi Hirano, and Masaru Sakurai. Super-resolution system for 4k-hdtv. In 2014 22nd International Conference on Pattern Recognition, pages 4453 – 4458. IEEE, 2014.
doi: https://doi.org/10.1109/icpr.2014.762

[10] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9199 – 9208, 2021.
doi: https://doi.org/10.1109/cvpr46437.2021.00908

[11] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image super-resolution via dual-state recurrent networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1654 – 1663, 2018.
doi: https://doi.org/10.1109/cvpr.2018.00178

[12] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1664 – 1673, 2018.
doi: https://doi.org/10.1109/cvpr.2018.00179

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770 – 778, 2016.
doi: https://doi.org/10.1109/cvpr.2016.90

[14] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1732 – 1741, 2019.
doi: https://doi.org/10.1109/cvpr.2019.00183

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132 – 7141, 2018.
doi: https://doi.org/10.1109/cvpr.2018.00745

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700 – 4708, 2017.
doi: https://doi.org/10.1109/cvpr.2017.243

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5197 – 5206, 2015.
doi: https://doi.org/10.1109/cvpr.2015.7299156

[18] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the 27th ACM International Conference on Multimedia, pages 2024 – 2032, 2019.
doi: https://doi.org/10.1145/3343031.3351084

[19] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. Advances in neural information processing systems, 29, 2016.

[20] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. IEEE Journal of solid-state circuits, 23(2):358 – 367, 1988.
doi: https://doi.org/10.1109/4.996

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply recursive convolutional network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1637 – 1645, 2016.
doi: https://doi.org/10.1109/cvpr.2016.181

[22] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral), June 2016.
doi: https://doi.org/10.1109/cvpr.2016.182

[23] Kwang In Kim and Younghee Kwon. Single-image super resolution using sparse regression and natural image prior. IEEE transactions on pattern analysis and machine intelligence, 32(6):1127 – 1133, 2010.
doi: https://doi.org/10.1109/tpami.2010.25

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
doi: https://doi.org/10.48550/arXiv.1412.6980

[25] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photorealistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681 – 4690, 2017.
doi: https://doi.org/10.1109/cvpr.2017.19

[26] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12321 – 12330, 2021.
doi: https://doi.org/10.1109/cvpr46437.2021.01214

[27] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In Proceedings of the European conference on computer vision (ECCV), pages 517 – 532, 2018.
doi: https://doi.org/10.1007/978-3-030-01237-3_32

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136 – 144, 2017.
doi: https://doi.org/10.1109/cvprw.2017.151

[29] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, volume 2, pages 416 – 423. IEEE, 2001.
doi: https://doi.org/10.1109/iccv.2001.937655

[30] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5690 – 5699, 2020.
doi: https://doi.org/10.1109/cvpr42600.2020.00573

[31] Karam Park, Jae Woong Soh, and Nam Ik Cho. Dynamic residual self-attention network for light weight single image super-resolution. IEEE Transactions on Multimedia, 2021.
doi: https://doi.org/10.1109/tmm.2021.3134172

[32] Sharon Peled and Yehezkel Yeshurun. Superresolution in mri: application to human white matter fiber tract visualization by diffusion tensor imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 45(1):29 – 35, 2001.
doi: https://doi.org/10.1002/1522-2594(200101)45:1<29::aid-mrm1005>3.0.co;2-z

[33] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In International conference on articulated motion and deformable objects, pages 175 – 184. Springer, 2016.
doi: https://doi.org/10.1007/978-3-319-41778-3_18

[34] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1874 – 1883, 2016.
doi: https://doi.org/10.1109/cvpr.2016.207

[35] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In International conference on medical image computing and computer-assisted intervention, pages 9 – 16. Springer, 2013.
doi: https://doi.org/10.1007/978-3-642-40760-4_2

[36] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3147 – 3155, 2017.
doi: https://doi.org/10.1109/cvpr.2017.298

[37] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE international conference on computer vision, pages 4539 – 4547, 2017.
doi: https://doi.org/10.1109/iccv.2017.486

[38] Matt W Thornton, Peter M Atkinson, and DA Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. International Journal of Remote Sensing, 27(3):473 – 491, 2006.
doi: https://doi.org/10.1080/01431160500207088

[39] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In

Proceedings of the IEEE international conference on computer vision, pages 1920 – 1927, 2013.
doi: https://doi.org/10.1109/iccv.2013.241

[40] Xuehui Wang, Qing Wang, Yuzhi Zhao, Junchi Yan, Lei Fan, and Long Chen. Lightweight single-image super-resolution network with attentive auxiliary feature learning. In Proceedings of the Asian Conference on Computer Vision, 2020.
doi: https://doi.org/10.1007/978-3-030-69532-3_17

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600 – 612, 2004.
doi: https://doi.org/10.1109/tip.2003.819861

[42] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. IEEE transactions on image processing, 19(8):2017 – 2028, 2010.
doi: https://doi.org/10.1109/tip.2010.2045707

[43] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In 2008 IEEE conference on computer vision and pattern recognition, pages 1 – 8. IEEE, 2008.
doi: https://doi.org/10.1109/cvpr.2008.4587647

[44] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. IEEE transactions on image processing, 19(11):2861 – 2873, 2010.
doi: https://doi.org/10.1109/TIP.2010.2050625

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In International conference on curves and surfaces, pages 711 – 730. Springer, 2010.
doi: https://doi.org/10.1007/978-3-642-27413-8_47

[46] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. Signal Processing, 90(3):848 – 859, 2010.
doi: https://doi.org/10.1016/j.sigpro.2009.09.002

[47] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 286 – 301, 2018.
doi: https://doi.org/10.1007/978-3-030-01234-2_18

[48] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082, 2019.
doi: https://doi.org/10.48550/arXiv.1903.10082

[49] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2472 – 2481, 2018.
doi: https://doi.org/10.1109/cvpr.2018.00262

[50] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6647 – 6656, 2021.
doi: https://doi.org/10.1109/cvpr46437.2021.00658

Introduction Authors

Karam Park

- 2018. 2 : B.S. in Department of ECE, Seoul National University
- ORCID : https://orcid.org/0000-0002-3612-0077
- Research interests : Image Processing based on Deep-learning

Nam Ik Cho

- Professor, Dept. of Electrical and Computer Engineering, Seoul National University
- 1986. 2 : Seoul National University, B.S. in Electrical Engineering (Dept. of Control and Instrumentation)
- 1988. 2 : Seoul National University, M.S. in Engineering
- 1992. 8 : Seoul National University, Ph.D. in Engineering
- ORCID : https://orcid.org/0000-0001-5297-4649
- Research interests : Digital Signal Processing, Image Processing, Adaptive Filtering, Computer Vision